Adapting the Burrows-Wheeler Transform (BWT) for Personalized DNA Analysis

Destiny Tudara
CS460

Introduction

- **Objective**: Adapt the BWT algorithm to enhance personalized DNA analysis.
- **Importance**: Empower users with private, efficient genetic data insights.
- **BWT Introduction**: Overview of BWT’s role in data compression and pattern matching.

Project Overview

Problem: Efficient, private analysis of raw DNA data from services like 23andMe.

Solution: Use BWT for data compression and analysis on personal computers.

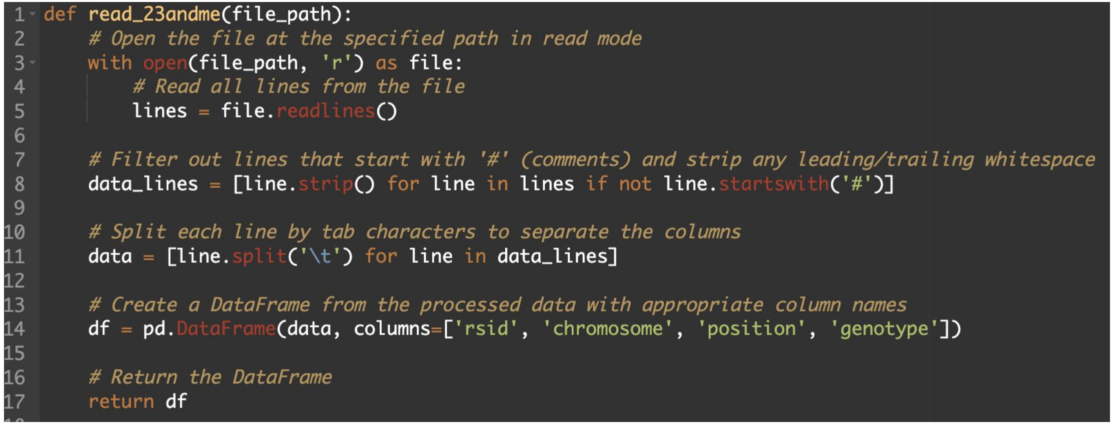
Benefits:

Increased accessibility of DNA analysis.

Ensures user privacy by performing analysis locally.

Improved efficiency in searching genetic data.

Methodology

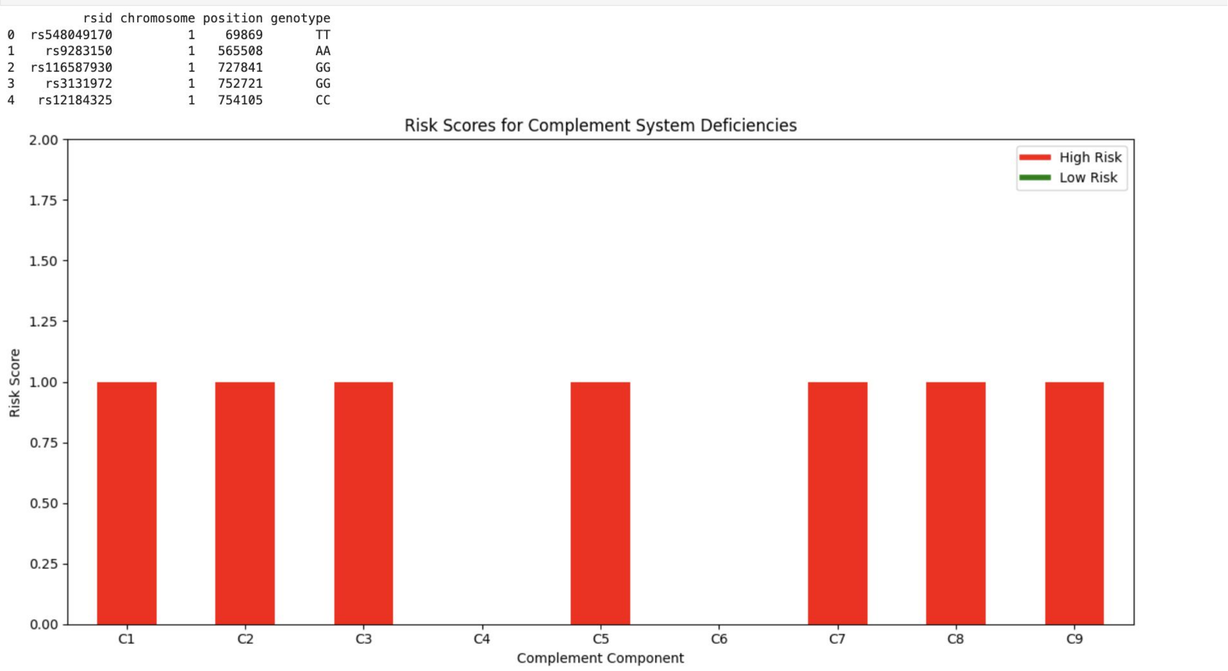


```python
def read_23andme(file_path):
    # Open the file at the specified path in read mode
    with open(file_path, 'r') as file:
        # Read all lines from the file
        lines = file.readlines()

    # Filter out lines that start with '#' (comments) and strip any leading/trailing whitespace
    data_lines = [line.strip() for line in lines if not line.startswith('#')]

    # Split each line by tab characters to separate the columns
    data = [line.split('\t') for line in data_lines]

    # Create a DataFrame from the processed data with appropriate column names
    df = pd.DataFrame(data, columns=['rsid', 'chromosome', 'position', 'genotype'])

    # Return the DataFrame
    return df
```

Content:

- **Data Processing**:
  - Reading raw DNA data files.
  - Extracting relevant genetic information (rsid, chromosome, position, genotype).
- **BWT Application**:
  - Transforming genetic data string using BWT.
  - Steps in BWT encoding and decoding.

Risk Assessment for Complement Deficiencies

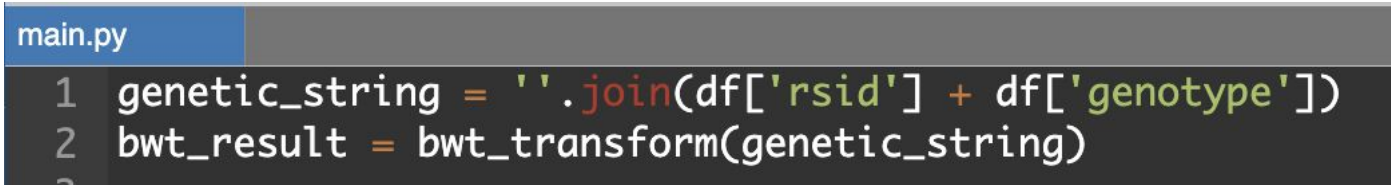


```
          rsid chromosome position genotype
0  rs548049170          1    69869       TT
1    rs9283150          1   565508       AA
2  rs116587930          1   727841       GG
3    rs3131972          1   752721       GG
4   rs12184325          1   754105       CC
```

Results

Content:

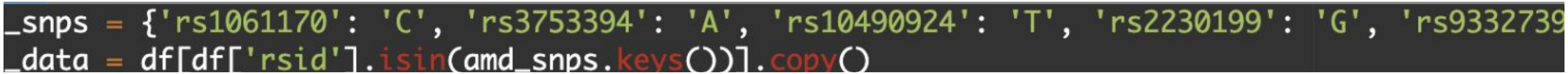
- **BWT Transformation**:
  - Example output of BWT transformation.



```
genetic_string = ''.join(df['rsid'] + df['genotype'])
bwt_result = bwt_transform(genetic_string)
```

Genetic Risk Analysis:

- AMD and Breast Cancer risk scores.
- Example risk score calculation.



```
_snps = {'rs1061170': 'C', 'rs3753394': 'A', 'rs10490924': 'T', 'rs2230199': 'G', 'rs9332739
_data = df[df['rsid'].isin(amd_snps.keys())].copy()
```

Conclusion and Future Work

Summary:

- Successful adaptation of BWT for DNA analysis.
- Enhanced privacy and efficiency.

Future Enhancements:

- Expanding analysis to include more genetic markers.
- Integration with other health data for comprehensive analysis.